Development and Application of a Supervised Pattern Recognition Algorithm for Identification of Fuel-Specific Emissions Profiles

Christos Stamatis[1] and Kelley C. Barsanti[1]

[1]Department of Chemical and Environmental Engineering and College of Engineering – Center for Environmental Research and Technology (CE-CERT), University of California, Riverside, Riverside, CA, USA

Correspondence: Kelley C. Barsanti (kbarsanti@engr.ucr.edu)

Abstract. Wildfires have increased in frequency and intensity in the western United States (U.S.) over the past decades. These trends are projected to continue, with negative consequences for air quality across the U.S. Wildfires emit large quantities of particles and gases that serve as air pollutants and their precursors, and can lead to severe air quality conditions over large spatial and long temporal scales. Characterization of the chemical constituents in smoke as a function of combustion conditions, fuel type and fuel component is an important step towards improving the prediction of air quality effects from fires and evaluating mitigation strategies. Building on the comprehensive characterization of gaseous non-methane organic compounds (NMOCs) identified in laboratory and field studies, a supervised pattern recognition algorithm was developed that successfully identified unique chemical speciation profiles among similar fuel types common in western coniferous forests. The algorithm was developed using laboratory data from single fuel species and tested on simplified synthetic fuel mixtures. The fuel types in the synthetic mixtures were differentiated, but as the relative mixing proportions became more similar, the differentiation became poorer. Using the results from the pattern recognition algorithm, a classification model based on linear discriminant analysis was trained to differentiate smoke samples based on the contribution(s) of dominant fuel type(s). The classification model was applied to field data and despite the complexity of contributing fuels, and the presence of fuels "unknown" to the classifier, the dominant sources/fuel types were identified. The pattern recognition and classification algorithms are a promising approach for identifying the types of fuels contributing to smoke samples and facilitating selection of appropriate chemical speciation profiles for predictive air quality modeling, using a highly reduced suite of measured NMOCs. Utility and performance of the pattern recognition and classification algorithms can be improved by expanding the training and test sets to include data from a broader range of single and mixed fuel types.

1 Introduction

Research has showed that the western United States[n] (U.S.) has seen an increase in the frequency and intensity of wildfires over the last three decades (Jaffe et al. (2020), Miller et al. (2009) and Dennison et al. (2014)), which is projected to continue (Westerling et al. (2006), Miller et al. (2009) and Dennison et al. (2014)). One of the consequences of wildfires is extremely poor air quality (McMeeking et al. (2005), McKenzie et al. (2006), Park et al. (2006), and Hu et al. (2018)). Emissions from wildfires include carbon monoxide (CO), carbon dioxide (CO2), and methane (CH4); several hundreds of gas-phase non-

methane organic compounds (NMOCs); and particulate matter (PM). While CO2 and CH4 are important greenhouse gases, NMOCs are of particular importance in the context of air quality because they serve as precursors to secondary air pollutants including photochemical ozone (O3) and secondary organic aerosol (SOA) (Andreae et al. (1988), Ward and Hardy (1991), Alvarado and Prinn (2009)). The latter of which, SOA, is a major constituent of atmospheric PM (Zhang et al. (2007)). In order to predict the air quality impacts of wildfires, differences in emissions and their effects on chemistry and pollutant formation must be represented in models (Kochanski et al. (2015), Pavlovic et al. (2016), Chen et al. (2019), Prichard et al. (2019), Jaffe et al. (2020)). Wildfire emissions are dependent on a number of factors such as combustion conditions (e.g., flaming vs. smoldering), fuel conditions (e.g., moisture content) and fuel type (e.g., species and component) (Goode et al. (2000), Urbanski (2013), Liu et al. (2017), Stockwell et al. (2014), Stockwell et al. (2015), Koss et al. (2018), Sekimoto et al. (2018), Hatch et al. (2019), Prichard et al. (2020)). Differences in these factors can affect the total amount~~flux~~[n] of emissions as well as the profile of emissions, i.e., the identities and quantities of individual chemical species~~constituents~~[n]. Permar et al. (2021) recently reported that combustion conditions, specifically modified combustion efficiency (MCE), explained approximately 70% of the variability in observed trace gas emissions from wildfires. Consistent with some existing modeling approaches, they suggested total NMOCs could be predicted using MCE, and the contribution of individual compounds determined using speciation profiles. Success of that approach requires knowledge of the relevant speciation profiles, and therefore contributing fuel types.

NMOC speciation profiles have been developed from both field and laboratory studies (Urbanski et al. (2008), Simpson et al. (2011), Urbanski (2014), Holder et al. (2017), Andreae (2019), and Prichard et al. (2020)). Laboratory studies offer some advantages over field studies in the context of controlling fuel species and fuel components; other variables, such as combustion conditions and fuel moisture, can be harder to control and can lead to differences in the identities and quantities of NMOCs emitted between laboratory and field studies (Yokelson et al. (2013), Stockwell et al. (2014), Liu et al. (2017), Sekimoto et al. (2018)). Yokelson et al. (2013) presented an inter-comparison of laboratory- and field-based emission factors (EFs), and approaches for using laboratory data to enhance the fundamental understanding of fire emissions coupled with field data to evaluate the representativeness of laboratory-based measurements. At that time, they noted that up to 70% of NMOCs remained unidentified for certain fuel types. More recently, due to the application of advanced instrumental techniques, there have been significant improvements in the identification and quantification of NMOCs emitted from fires, particularly in laboratory studies (Stockwell et al. (2014), Stockwell et al. (2015), Hatch et al. (2017), Koss et al. (2018)). For example, Stockwell et al. (2015) detected approximately 80–96 % of the total emitted NMOC mass in experiments during the 2012 fourth Fire Lab at Missoula Experiment (FLAME-4); and Hatch et al. (2019) identified more than 500 individual NMOCs during FLAME-4. The relatively rapid expansion in available NMOC data provides opportunities for developing more detailed speciation profiles (in which a higher fraction of the detected mass is assigned to unique compounds or formulas) and for applying statistical data analysis methods, facilitating the identification of unique sets of compounds that allow differentiation of fuel type(s) and estimation of their contributions to smoke samples.

Existing approaches for identifying the contribution of fuel types to smoke include land cover databases or fuel loading models coupled with fuel consumption models (e.g. FOFEM, Keane and Lutes (2018); and CONSUME, Ottmar (2009)), and

the use of marker compounds. One of the limitations of land cover databases or fuel loading models is that they are difficult to update frequently enough to reflect changes in ecosystems (Reeves et al. (2009), Vogelmann et al. (2011), Nelson et al. (2013) and Lindaas et al. (2021)). Marker compounds are emitted in relatively high abundances and can be used to differentiate fuels by component or fuel layer and in some case by species. For example, Wan et al. (2019) showed that *p*-hydroxybenzoic acid was emitted from combustion of herbaceous plants, while vanillic acid was emitted from combustion of softwoods and hardwoods. It has also been shown that syringic acid is associated with hardwood combustion (Simoneit (2002) and Zangrando et al. (2013)), and dehydroabietic acid with conifers (Fu et al. (2009)). Zhang et al. (2021) found that the benzene to toluene ratio in smoke from sugarcane leaves was different than the ratio in smoke from sesame stalk, demonstrating differences among agricultural fuels. In measurements of western forests and shrublands, Jen et al. (2018) showed that hydroquinone was a good marker for manzanita combustion. One of the limitations of using marker compounds to identify fuel types is the lack of specificity, i.e., marker compounds have not been identified that enable identification of a large number of fuel species or closely related fuel species.

In this work a method is presented for identifying fuel types from measured NMOCs in smoke samples. To overcome some of the existing limitations in identifying the contribution of specific fuel types to smoke, pattern recognition (PR) and classification algorithms were developed using data obtained during two laboratory campaigns in 2012 and 2016, and applied to data obtained during a field study in 2017. Machine learning techniques have been applied for source identification in other disciplines. For example, Welke et al. (2013) and Ziółkowska et al. (2016) used principal component analysis (PCA) and linear discriminant analysis (LDA) to differentiate and classify wine varietals based on specific compounds present in wine samples. Johnson and Synovec (2002) used PCA and analysis of variance to select marker compounds in gasoline fuel blends and PR to differentiate the blends. In this work, the large data sets generated during FLAME-4 and the Fire Influence on Regional to Global Environments Experiment (FIREX) 2016 Fire Lab campaigns were leveraged to develop a source identification method using fuel-specific NMOC profiles. The PR algorithm performs an automated selection of compounds that differentiate sources (in this case, fuels) based on measured NMOCs. The classification algorithm then uses the source profiles to identify source contributions to specific samples. The data used to train and test the algorithm are introduced in section 2. The algorithm development, implementation and testing are presented in sections 2 and 3. The application to field data is presented in section 3, and general conclusions and implications are presented in section 4.

2 Data and Methods

2.1 Data

The NMOC data used in this study were acquired from a variety of fuel types burned in laboratory and field settings during three campaigns: 1) FLAME-4 laboratory campaign in 2012 (FLAME-4 FL12), 2) FIREX laboratory campaign in 2016 (FIREX FL16), and 3) Blodgett Forest Research Station (BFRS) prescribed burns in 2017; both laboratory campaigns took place at the U.S. Forest Service Fire Science Laboratory (FSL). Details of the facilities, sample collection and data analysis have been discussed in previous publications (Stockwell et al. (2014), Hatch et al. (2015), Hatch et al. (2019), Selimovic et al. (2018)).

Briefly, during FLAME-4 FL12 and FIREX FL16, a broad variety of biomass fuels were burned (Stockwell et al. (2014), Selimovic et al. (2018)), including conifers and shrubs (Table 1); 80 samples were collected from both room and stack burns as described in Stockwell et al. (2014) and Selimovic et al. (2018). During the BFRS study, a total of 28 samples (Hatch et al. (2019)) were collected from a utility task vehicle parked downwind from three separate prescribed burn plots that had different fuel distributions (see Table1 and Supplementary Information (SI) Figs. S2-S4 in Hatch et al. (2019)). All NMOC samples were collected using dual bed stainless steel sorbent tubes and were analyzed using an automated thermal desorption unit coupled to a two dimensional gas chromatograph with a time-of-flight mass spectrometer (GC × GC-TOFMS). The raw chromatograms were processed using the commercially available software Chromatof (Leco Corp., St. Joseph, MI). The measured mixing ratios were used to calculate normalized excess mixing ratios (NEMRs) versus CO, $\Delta X/\Delta\mathrm{CO}$ (Yokelson et al. (1999)), in which delta represents excess over background. The calculated NEMRs of monoterpenoids ($C_{10}H_{16}$ and $C_{10}H_{16}O$) were used as the starting point for this analysis based on Hatch et al. (2018) and Hatch et al. (2019), and the emission profile analysis presented in the SI section 1[n]. Hatch et al. (2019) demonstrated that the variability in NMOC composition could not be attributed entirely to MCE, and that chemical speciation was highly correlated among some fuel types across a range of MCE values, particularly conifers; within conifers, clear differences in monoterpenoid emissions were observed as a function of fuel species.

2.2 Pattern recognition algorithm

A four-step PR algorithm (Fig. 1) was developed to select a subset of compounds that captured the variance between fuel types and then use the selected compounds to differentiate fuel types based on NMOC speciation profiles; the algorithm steps are: 1) data preprocessing, 2) analysis of variance (ANOVA)[n], 3) principal component analysis (PCA)[n] and 4) and *k*-means clustering. The algorithm was implemented using the Python package scikit-learn (Pedregosa et al. (2011)). The algorithm components[n] are explained in sections 2.2.1-2.2.4. Implementation of the algorithm is explained in section 2.3.[n].

2.2.1 Preprocessing and analysis of variance (ANOVA) ~~feature selection~~[n]

Data preprocessing (step 1) is performed to handle any missing values in the samples. Approaches for handling missing values are specific to the type(s) of data and the reason(s) for missing values (Dong and Peng (2013); McNeish (2017)). In this data set, missing values largely were a result of compounds being below the detection limit or having negative values after background correction. During preprocessing, for every feature (i.e., compound) the percentage by number of missing values across all samples was calculated. For any given compound, if the percentage of missing values was less than 30% then the missing values were replaced with zeros. If the percentage was more than 30% then the compound was removed from the data set. The 30% threshold is supported by published statistical methods guides, including Dong and Peng (2013) and Jakobsen et al. (2017), that suggested a threshold range of 10% - 40% prior to replacement.[n]

Samples also were evaluated and filtered for missingness, using two criteria. For criterion one, only fuel types that had more than 30% (by number) of the 93 monoterpenoids above background levels were selected. Since the PR algorithm was based on monoterpenoids, samples with few to no detected monoterpenoids would reduce the ability of the algorithm to differentiate between fuel types and therefore reduce the overall efficiency. For criterion two, only fuel types that had three or more samples

Table 1. Fuels burned and smoke analyzed from FLAME-4 2012 laboratory fires, FIREX 2016 laboratory fires and Blodgett Forest Research Station prescribed burns.

Fuel Family	Fuel Type	FLAME-4 FL12 (lab)	FIREX FL16 (lab)	BFRS
Conifers				
	Ponderosa pine	x	x	x
	Lodgepole pine		x	
	Engelmann spruce		x	
	Black spruce	x		
	Douglas fir		x	
	Subalpine fir	x	x	
	White fir			x
	Juniper		x	
	Loblolly pine		x	
	Sugar pine			x
	Jeffrey pine		x	
	Incense cedar			x
Shrubs				
	Chamise		x	
	Manzanita		x	
	Sagebrush		x	
	Snowbrush ceanothus		x	
Miscellaneous				
	California black oak	x		
	Tanoak			x
	Excelsior		x	
	Yak dung		x	
	Peat	x	x	
	Rice straw	x	x	
	Bear grass		x	
	Untreated lumber	x		

were retained. For fuels with too few samples, the necessary statistical properties can not be calculated. In addition, the second criterion ensured that the remaining fuel types had an approximately equal number of samples in the analysis.[n]

Feature selection (step 2) reduces the number of variables by selecting only those that are informative. In this work, an analysis of variance (ANOVA)-based feature selection method, similar to Johnson and Synovec (2002) and Welke et al. (2013),

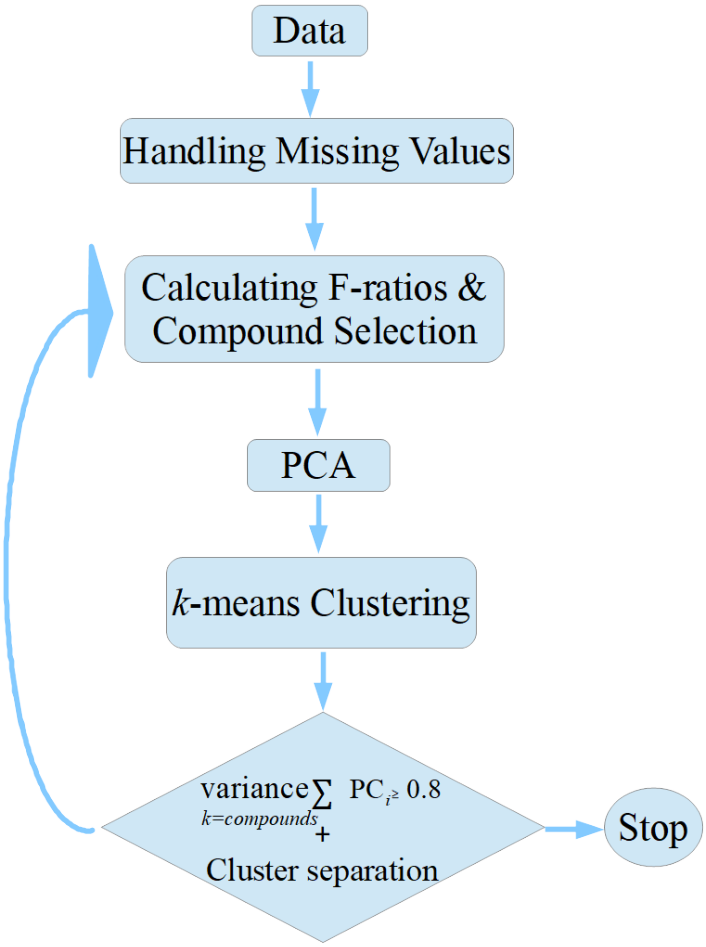

Figure 1. Pattern recognition algorithm flowchart.

was used to further filter the compounds retained in step 1. In this application, each detected compound was treated as an ANOVA-type problem with N samples in k classes (fuel types) to determine whether or not a compound could separate the different fuel types. For each compound, the ratio of class-to-class variance to within-class variance was calculated using Eq. 1, also known as the Fisher ratio (F-ratio):[n]

$$F_{ratio} = \frac{V_b}{V_w} \quad (1)$$

[n]where the nominator (V_b) corresponds to the between-class sum of squares and the denominator (V_w) to the within-class sum of squares. The magnitude of the F-ratio is an indication of class separation. Following the F-ratio calculation the compounds were ranked in an ascending order based on their F-ratios values. Further details regarding the F-ratio calculation can be found in SI section 2.[n]

2.2.2 Principal component analysis (PCA) ~~Principal component analysis and k-means clustering~~[n]

PCA (step 3), as described in Abdi and Williams (2010), is a dimensionality reduction technique that is used to project high dimensional data into a lower dimensional space along the direction(s) of maximum variance in the data. ~~k-means (Jain (2010)) (step 4) is a popular clustering algorithm that finds clusters in a n-dimensional space (Jolliffe (2002) and Abdi and Williams (2010)). In this study PCA was used to compress the information carried by the selected compounds to a lower dimensional space and k-means clustering was used to find formed clusters after the application of PCA.~~[d] In PCA the original variables are transformed into a new coordinate system. The new coordinate system is formed using the calculated principal components (PCs) which serve as the new directions/axes. Each PC is a linear combination of the original variables and weights (also called loadings), as shown in Eq. 2:[n]

$$PC_i = \sum_{ij=1}^{m} w_{ij} x_j \quad (2)$$

[n] where *i* and *j* are indices corresponding to the number of PCs and number of original variables, respectively. Each PC accounts for a specific amount of variance of the observed data, with the first accounting for the greatest amount and each succeeding component an amount less than the previous. The PCs can be calculated either by eigenvalue decomposition (EVD) on the covariance matrix of the original data or singular value decomposition (SVD) on the data matrix. In this study SVD was used; the implementation of SVD is further described in SI section 3. Regardless of the method being used to calculate the PCs, because PCA is trying to find new axes along the direction(s) of maximum variance, if the variables are not on the same scale or if there is a large difference in the magnitudes, the results can be biased in favour of the variable(s) that vary on a larger scale (Gewers et al. (2021), Lever et al. (2017)). For this reason standardization of the data might be necessary prior to PCA, in which the data are mean centered and each variable is divided by the standard deviation of that variable. While standardization can help alleviate the scaling problem it should not be a default practice as it can magnify the effect of outliers in the data (Gewers et al. (2021)). If the variability of a feature is a consequence of intrinsic variability in the analytical method (e.g., experimental error or noise in the data), then standardization may erroneously emphasize that in the PCA results. In such cases, either the noise should be reduced by some means or standardization should be avoided. In this study, the selection of PCA for dimensionality reduction was based on three previous studies that included PCA in pattern recognition analysis of chromatographic data (Welke et al. (2013), Johnson and Synovec (2002), Ziółkowska et al. (2016)).[n]

2.2.3 *k*-means clustering[n]

k-means (step 4) is a popular clustering algorithm that finds clusters in an *n*-dimensional space (Jolliffe (2002)). Given a set of observations (x_1, x_2 , x_3, ... x_n) where each observation is an *d*-dimensional real vector, *k*-means tries to partition the *n* observations into $k \leq n$ sets S = {S_1, S_2 , S_3 , ..., S_k}. Mathematically, *k*-means clustering minimizes within cluster variances, or squared Euclidean distances (Jain (2010)), as shown in Eq. 3:[n]

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S} ||x - \mu_i||^2 \quad (3)$$

[n] where μ_i is the mean of points in S_i. In this study Elkan's algorithm (Elkan (2003)) was used to solve Eq 3. *k*-means clustering was used to find formed clusters after the application of PCA. The inputs for the clustering analysis were the retained PCs. *k*-means was chosen over other clustering algorithms because of its simplicity and the absence of highly anisotropically distributed clusters (irregular shapes); unequal variance among the clusters and unevenly sized clusters can cause problems with *k*-means (Jain (2010))[n].

2.2.4 Determining the number of principal components and clusters[n]

Three metrics were considered for determining the number of components retained after PCA: 1) explained variance based on the number of retained components, 2) the Kaiser criterion, and 3) the scree test. The explained variance % (SI Eq. 8) evaluates how much of the original variance in the data is explained by each additional component retained. For this metric there is no strict threshold; 80% was chosen for this study. The Kaiser criterion suggests rejecting all factors that have eigenvalues less than 1. The scree test requires plotting the eigenvalues as a function of component number; an inflection point indicates the maximum number of usable components. The default method used in this study was the explained variance but all three approaches were evaluated and provided similar results.[n]

k-means requires that the target number of clusters are provided in advance. This is challenging when the number of clusters is not known. In this study, the elbow plot method was used to determine the number of clusters (Fig. 5). The elbow plot method requires running the *k*-means multiple times using a different number of clusters each time. For each run the total within-sum of squares (TWSS) was calculated according to Eq. 4:[n]

$$d = \sum_{k=1}^{k_{max}} \sum_{j=1}^{n} \sqrt{(q_j - c_k)^2} \quad (4)$$

[n] where q and c are multidimensional vectors with the coordinates for each centroid and sample, respectively; n is the total number of samples and k_{max} is the total number of pre-selected clusters. TWSS is a measure of variability for the observations within a cluster. The smaller the TWSS for a number of clusters the better the clustering. TWSS is plotted against the number of clusters. As with the scree test the inflection point in the plot signifies the optimum number of clusters.[n]

2.3 Running the pattern recognition algorithm[n]

Implementation of the PR algorithm proceeds through five steps. In step one, the compounds in the data set are preprocessed to replace missing values and discard samples that might be problematic (section 2.2.1). In step two, for each retained compound the *F*-ratio is calculated. In step three, in an iterative fashion, PCA and *k*-means clustering are performed on the *m* highest ranked compounds and the separation of the classes in the samples (fuel types) is evaluated as a function of the number of compounds and the minimum number of PCs that achieve the 80% variance threshold. In step four, if the separation is not adequate then the number of compounds is increased or decreased and the run is repeated. In step five, once the class separation no longer improves or starts degrading with the addition or removal of compounds (step 4), more PCs are retained and different combinations of PCs are tested. While increasing the number of components will lead to better separation of the fuel types, it can also lead to overfitting. In this study the algorithm was optimized for sample separation as a function of the compounds selected and pairing of the PCs. The effects of overfitting were not considered in the optimization due to limitations in the sample size that prevented the use of known evaluation methods.[n]

2.4 Classification

To test the applicability of the PR algorithm results for field samples, a classification algorithm, LDA (Hastie et al. (2009)), was applied (see SI section 5 for implementation details). LDA is a supervised learning method that is similar to PCA. Both LDA and PCA are linear transformation techniques; LDA is supervised, whereas PCA is unsupervised and ignores class labels. While PCA tries to find a subspace of features in order to maximize variance among samples, LDA attempts to find a feature subspace that maximizes class separability. The inputs for the LDA training were the selected PCs from the PR analysis (independent variables) and the fuel types (response variable/class) (see section 4 in SI). The output of LDA is a probability score for every sample that is being tested for its likelihood to belong to a particular fuel type, calculated as follows (Eq. 5):[n]

$$\log P(y=k|x) = -\frac{1}{2}(x-\mu_k)^t \Sigma^{-1}(x-\mu_k) + logP(y=k) + C_{st} \quad (5)$$

[n] where k is the class of sample x, μ is the vector of the means for each class based on the selected features, Σ is the common covariance matrix for the three classes in the training set and C_{st} is a term that contains constants from the multivariate Gaussian distribution. LDA was chosen as the classification method in this study because of its closed-form solution that does not require any hyperparameter tuning. Methods that require hyperparameter tuning (e.g., k-Nearest Neighbours) are not appropriate for this application because: 1) the training data set included laboratory data only while the test set included field data and 2) the field data were not labeled.[n]

3 ~~Algorithm Implementation,~~ [d]Results and Discussion

3.1 Sample and fuel type selection for pattern recognition and classification

The PR algorithm was applied to the FIREX FL16 data set to identify a group of marker compounds that could be used to differentiate fuel types. Classification was then performed using the FIREX FL16 data as the training set, and BFRS data as the testing set. The selection of the training and testing sets was based on the size of each data set; the FIREX FL16 data

Table 2. Data sets used for developing the pattern recognition algorithm and testing and training classification algorithm.

Data Set	Pattern Recognition	Training Set	Testing Set
FLAME-4 FL12			x
BFRS			x
FIREX FL16	x	x	
Synthetic data	x		x

set had 74 samples and the BFRS data set had 29 samples. A larger training set ensured more statistically robust parameters for the LDA model. Because BFRS samples were generated from combustion of complex fuels, a synthetic data set was generated to test the performance of the PR and classification algorithms on mixed fuel samples prior to application on the BFRS data. The synthetic samples were created by taking the average value of each selected NMOC from each fuel type and linearly combining them using specific percentages for two fuels (e.g., 60/40 and 90/10).[n] The five synthetic mixtures had the following~~Five synthetic mixtures were generated with the~~[n] compositions: 60% pine/40% spruce, 60% fir/40% spruce, 60% pine/40% fir, 90% pine/10% spruce and 90% fir/10% spruce. The FLAME-4 FL12 data were used as an independent data set to test the response of the classification algorithm to fuel types that were not included in the training set. The use of each data set in the PR and classification algorithms is summarized in Table 2.

~~Two selection criteria were applied to the training set to ensure that standard deviations and averages could be computed, which are central features of the PR algorithm. First, only fuel types that had more than 30% (by number) of the 93 monoterpenoids above background levels were selected. Since the PR algorithm was based on monoterpenoids, samples with little to no detected monoterpenoids would reduce the ability of the algorithm to differentiate between fuel types and therefore reduce the overall efficiency. Second, only fuel types that had three or more samples were retained.~~[d] Before applying the PR algorithm the data were processed as described in section 2.2.1.[n] Application of the preprocessing criteria reduced the number of samples from a total of 74 to 39 and the number of fuel species from 18 to five: pines (ponderosa pine and lodgepole pine), firs (Douglas fir and subalpine fir) and spruce (Engelmann spruce). During the FIREX FL16 study, different fuel components were burned such as canopy, rotten log, composite, litter and duff. While differences in component emissions may be important for differentiating prescribed burn and wildfire smoke samples, for this application, 32 composite and canopy samples, and seven litter and duff samples were retained based on the selection criteria (section 2.2.1).

3.2 Pattern recognition

3.2.1 Feature selection

Feature selection was performed and evaluated using two approaches: 1) manual selection, where the compounds were filtered based on a single criterion, whether a compound was present in more than three fuel species;[n] and 2) automated selection using the the PR algorithm (section 2.3) ~~on Fisher ratios (F-ratios)~~[n]. ~~The percentage of variance explained using the first two PCs~~

was used as the metric to evaluate the quality of feature selection using the two approaches (see scree plot method section 3.2.2) following application of PCA (Fig. 2). For automated selection, the F-ratios were calculated for every compound in the data set using Eq. 1. In Eq. 1 the nominator (V_b) corresponds to the between-class sum of squares and the denominator (V_w) to the within-class sum of squares.[d] The selection approaches were compared using the metrics introduced in section 2.2.4 and visualizing the emission profiles of the selected compounds.[n] The manual approach resulted in the selection of the following nine (out of 93) compounds: a-pinene, limonene, 3-carene, b-myrcene, camphene, p-cymene, bornyl acetate, b-phellandrene and tricyclene. The automated approach resulted in selection of the following five compounds: tricyclene, camphene, b-pinene, 3-carene and bornyl acetate. Figure 2 shows the improved performance of automated feature selection over manual feature selection, based on the single highest explained variance across PCs 1-5.[d] The PR algorithm was run again such that the number of selected compounds (nine) were the same between the manual and automated selection methods. The cumulative explained variance plots for manual and automated selection of compounds are shown in Fig. 2. Given the 80% threshold, three components were required with manual and automated selection of nine compounds, and two components with automated selection of five compounds. Using the Kaiser criterion, three components were necessary for an effective separation with the manual (Fig. S2) and automated (Fig. S3) selection of nine compounds, and only one component with the automated selection of five compounds (Fig. S4). Using the scree test, five components were needed with the manual selection of nine compounds (Fig. S2) and two components with the automated selection of five compounds (Fig. S4). Regardless of the metric used to evaluate the quality of the feature selection, the automated selection of five compounds always resulted in the lowest number of PCs, one or two, to meet the criteria specific to that metric. A comparison of the number of PCs required relative to the number of original dimensions is presented in the SI (section 6), which is another measure of the effectiveness of the feature selection.[n]

To make the feature selection results more intuitive, the normalized emission ratio profiles (ratio of the compound ER to the sum of ERs for the selected compounds) as a function of fuel species are shown for manual selection (Fig. 3), and automated selection of five (Fig. 4) and nine (Fig. S5) compounds. Emerging patterns can be seen in the resulting profiles between and within the fuel types. The emission profiles from the automated selection, for both five and nine compounds, provide more distinct profiles between fuel types, and more consistent profiles within types, than the profiles from manual selection. For example, with manual selection (Fig. 3) the normalized emission ratio profiles show that the the relative abundances of alpha-pinene (black) and d-limonene (dark brown) are more similar between ponderosa pine and Engelmann spruce than they are for the two pines, and the two pine species have dissimilar relative abundances of 3-carene (yellow) and b-phellandrene (tan). However, these differences within pines and similarities between ponderosa pine and spruce disappear with the automated selection, particularly with five compounds. The consistency of profiles within the fuel types is important because it allows for fuel separation and classification when new samples are provided. Given the more effective dimensionality reduction of the automated feature selection, as well as the more consistent emission profiles, the manual feature selection will be not be further discussed.[n] The automated selection with five compounds results in more distinct and consistent profiles for each fuel family, which translates to a higher potential for separation (greater explained variance) in the PCA space. The five compounds selected with the automated approach were thus used for the PR analysis.[d]

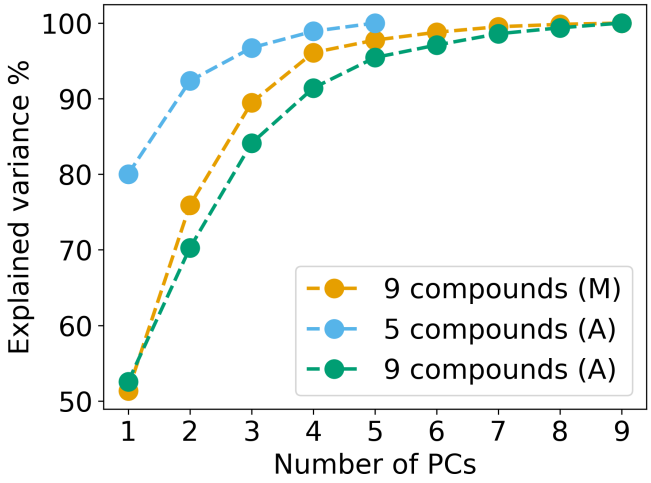

Figure 2. Cumulative explained variance for automated (blue and green) and manual (orange) compound selection.

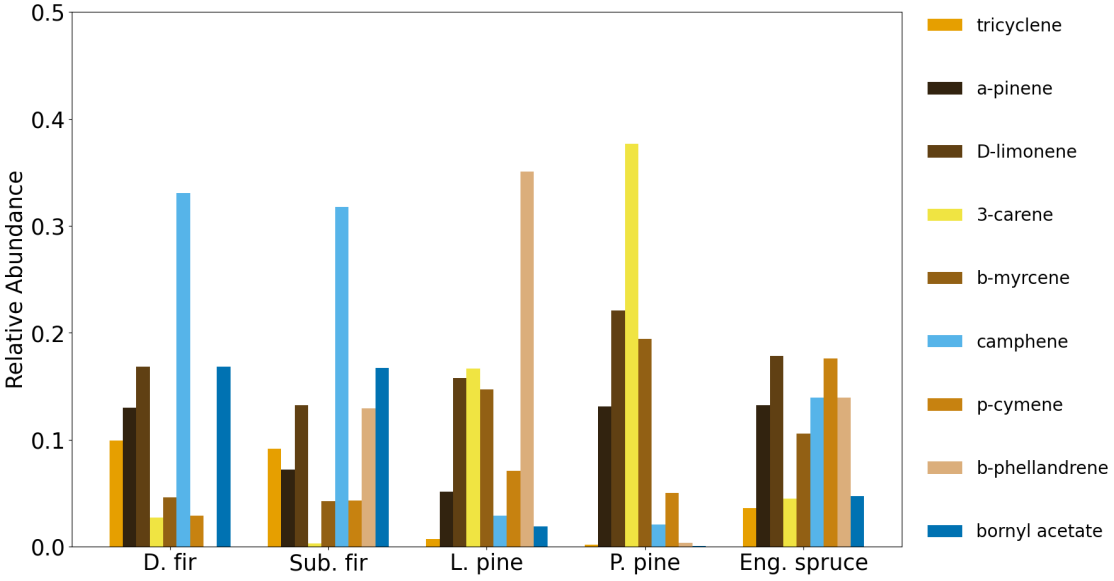

Figure 3. Normalized emission ratio profiles for: Douglas fir, subalpine fir, lodgepole pine, ponderosa pine and Engelmann spruce based on manual compound selection.

3.2.2 PCA and *k*-means clustering

285 ~~Following data preprocessing and feature selection, PCA was performed on the reduced data set. To determine the number of PCs to be retained, a scree test using a modified version of the Kaiser criterion (Jolliffe (2002)) was performed. In the~~

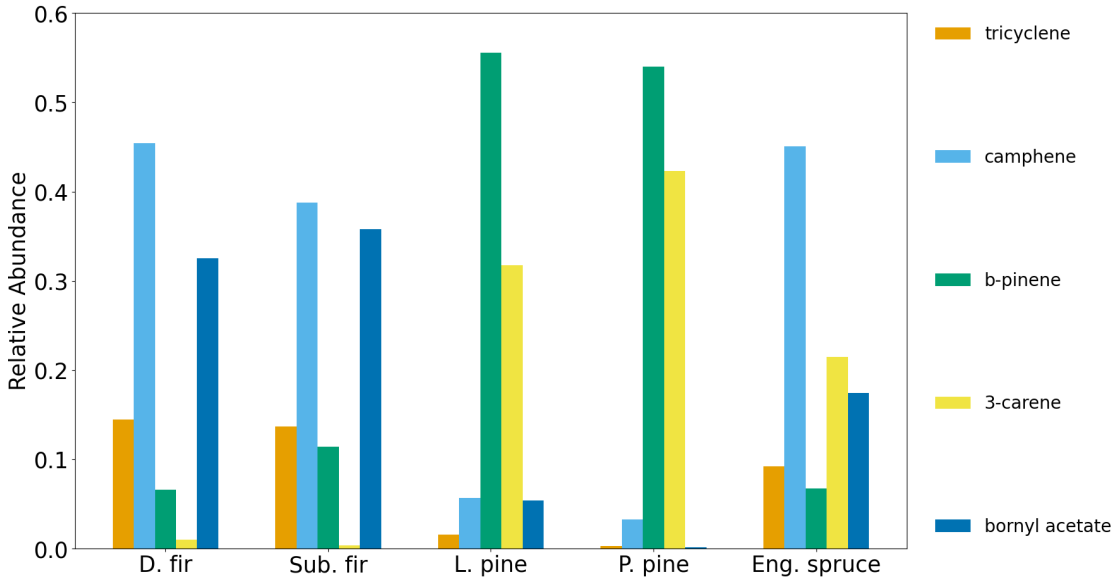

Figure 4. Normalized emission ratio profiles for: Douglas fir, subalpine fir, lodgepole pine, ponderosa pine and Engelmann spruce based on automated compound selection.

~~scree test, the component number is determined by plotting the acquired eigenvalue (or explained variance) as a function of component number. A steep decrease or inflection point indicates the number of usable components. In this study the normalized eigenvalues were calculated using Eq. 2 and plotted against the number of components (Fig. 2).~~[d]

$$\frac{I_j}{\sum_{j=1}^{p} I_j} \quad (2)^{\text{d}}$$

~~where I_j corresponds to the eigenvalue for PC$_j$. Figure 2 shows that with automated feature selection two PCs (PC1 and PC2) were adequate to explain 92% of the variance in the data set. The scores from these retained PCs (PC1 and PC2) were then used as input for k-means clustering. To determine the optimal number of clusters, the elbow plot method was used. The elbow plot is a graphical method in which the cumulative distance is calculated for all points/samples from their respective centroids and then plotted against the number of clusters. In this study, Euclidean distance was used (Eq. 3):~~[d]

$$d = \sum_{k=1}^{k_{max}} \sum_{j=1}^{n} \sqrt{(q_j - c_k)^2} \quad (3)^{\text{d}}$$

~~where q and c are multidimensional vectors with the coordinates for each centroid and sample, respectively. The values for each sample are the respective scores from the selected PCs. The indices j and k correspond to centroid number and sample number, respectively. The optimum number of clusters is found after a steep decrease in the total Euclidian distance, followed by trivial changes, with increasing number of clusters. A steep decrease signifies good clustering performance (dense clusters),~~

~~while a trivial change in the Euclidian distance shows that the cluster centroids do not change substantially from their previous positions. The elbow plot (Fig. 5) shows that the algorithm identified four clusters as the optimum number.~~[d] Following data preprocessing and feature selection, PCA was performed on the reduced data set (five compounds) followed by k-means clustering on two retained components (based on the explained variance metric, see Fig. 2). For *k*-means the number of clusters was determined using the elbow plot method (section 2.2.4). In Fig. 5 TWSS is shown as a function of the number of clusters; the last inflection point occurs around four clusters (dark blue marker). The coupled PCA and *k*-means results are shown in Fig. 6. The clusters identified by the algorithm are differentiated by marker color while the fuel types are differentiated by marker shape. Cluster one included 15 out of 16 pine samples and one overlapping fir sample. Cluster two included 11 out of 13 fir samples and four overlapping spruce samples. Clusters three and four included the remaining six spruce samples, one overlapping pine sample and one overlapping fir sample. Generally the algorithm resulted in adequate separation between firs and pines; but poorer separation for spruce, for which four of ten samples overlapped with another fuel family. Adding four more compounds reduced the explained variance from PC1 and PC2 from 92% to less than 70% and resulted in only minor improvement in cluster separation (Fig. S6). The difficulty that the algorithm encounters separating spruce effectively can be explained using the elbow plot (Fig. 5). The *k*-means algorithm identified four clusters as the optimum number, but the steep decrease in the total TWSS~~total Euclidean distance~~[n] actually occurs between one and two total clusters. The TWSS ~~total Euclidean distance~~[n] decreases further between two and four but to a lesser extent (shallower slope). The lesser decrease between two and four clusters indicates that the clustering algorithm had difficulty identifying clusters in the PCA space, which is then apparent in Fig. 6. From the normalized emission ratio profiles (Fig. 4), it can be seen that the spruce and fir samples have similar normalized emission ratios for tricyclene, camphene and b-pinene. This limits the ability to fully separate spruce and firs in the PCA space.[n]

3.2.3 Beyond principal components one and two

Thus far, only the first two PCs (from a total of five) were used in the analysis since they explained 92% of the variance in the data set. Another 8% was shared between PCs three and four, which could potentially provide better separation for the spruce samples. After testing PCs one with three and one with four, the combination of one and four resulted in better performance of the PR algorithm. Though the optimal number of clusters for the PC1 and PC4 pair (Fig. 7) was the same as with the PC1 and PC2 pair, it was found at a lower TWSS~~total Euclidean distance~~[n], which is indicative of more dense clustering. Figure 8 shows the results for the PC1 and PC4 pair. Cluster one included 13 out of 16 pine samples and one overlapping fir sample. Cluster two included 11 out of 13 fir samples and only one overlapping spruce sample. Clusters three and four included nine out of 10 spruce samples, one overlapping fir sample and three overlapping pine samples. With the PC1 and PC4 pair, spruce samples had 30% less overlap with firs (Fig. 9), with only moderate losses in the separation between spruce and pines. These results demonstrate the ability of the PR algorithm to separate firs, pines and spruce in the smoke samples, with only two PCs accounting for most of the variance in the data set (PC1 and PC4 about 82%).

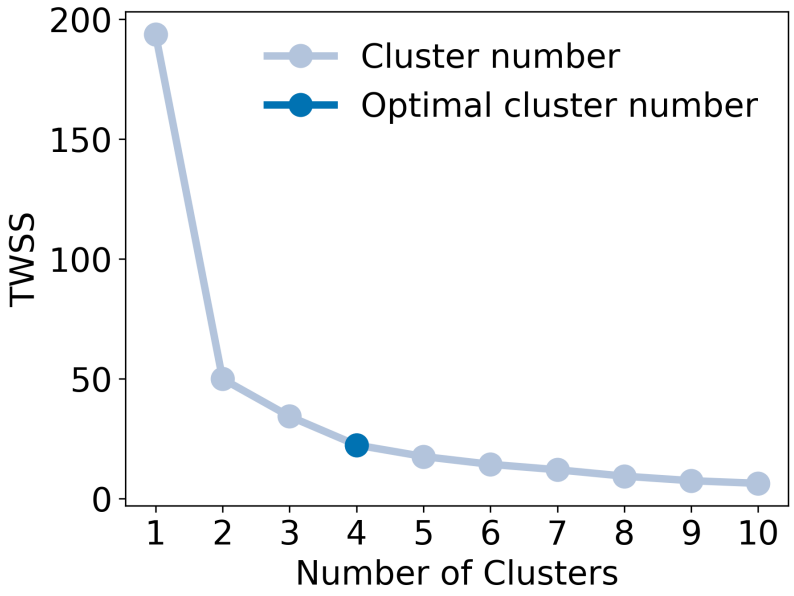

Figure 5. Elbow plot for *k*-means clustering with automated compound selection for the PC1 and PC2 pair. The dark blue marker indicates the optimum number of clusters.

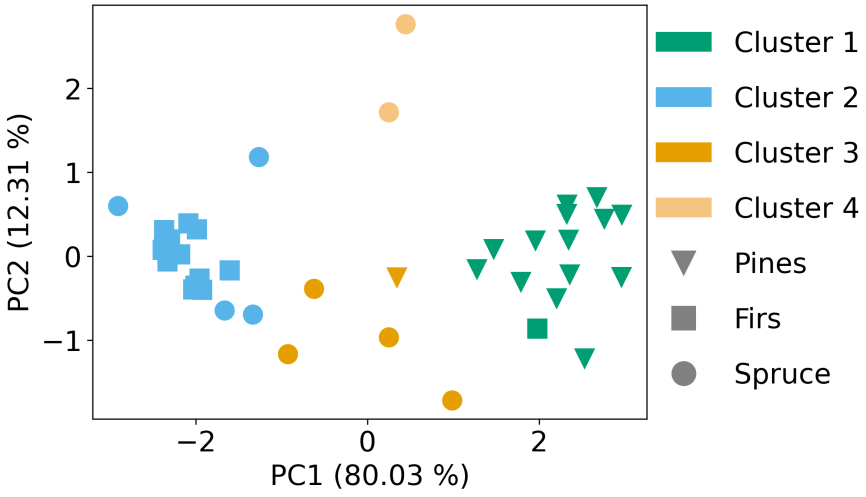

Figure 6. PCA coupled with *k*-means clustering results for the PC1 and PC2 pair.

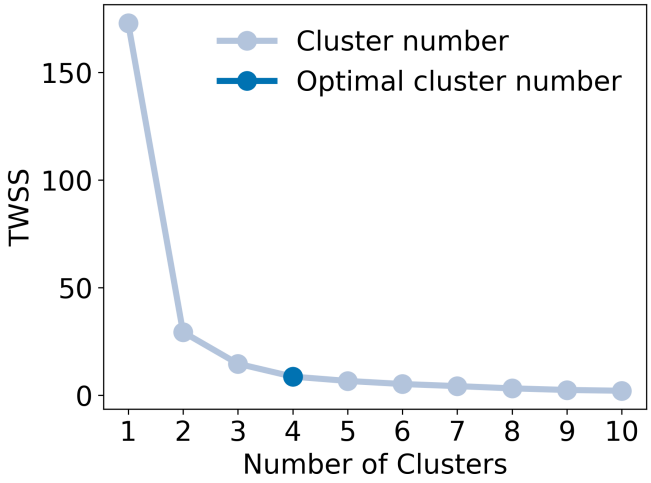

Figure 7. Elbow plot for *k*-means clustering with automated compound selection for the PC1 and PC4 pair. The dark blue marker indicates the optimum number of clusters.

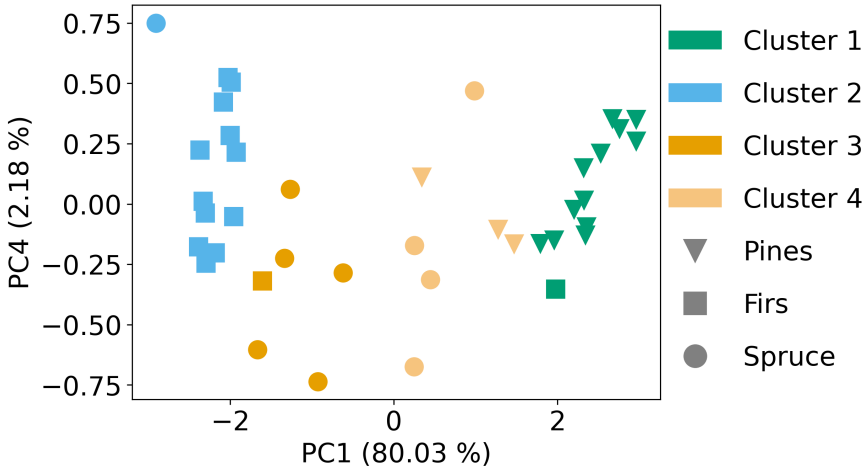

Figure 8. PCA coupled with *k*-means clustering results for the PC1 and PC4 pair.

3.2.4 Mixed samples

The PR algorithm selected compounds that separated single-fuel smoke samples by the contribution of fuels categorized into three types (firs, pines and spruce). Before testing the algorithm on complex smoke samples the algorithm was tested on synthetic fuel mixtures described in section 3.1. From the three 60/40 samples only the fir/spruce synthetic mixture was clustered with the dominant fuel family (fir). The pine/spruce and pine/fir synthetic mixtures were clustered with spruce clusters one and

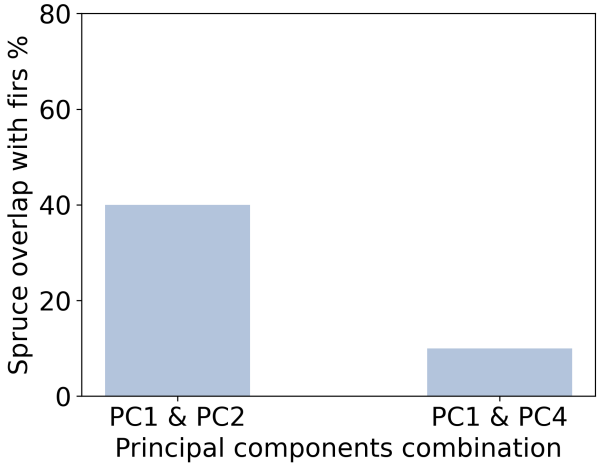

Figure 9. Spruce overlap with firs based on the PC combination used.

three, respectively (Fig. 10). The clustering of the pine/spruce synthetic mixture as spruce was marginal in the PCA space and was due to the scatter of the spruce samples rather than the similarity of the synthetic mixture with spruce. The clustering of the pine/fir mixture with spruce is more intuitive after comparing the normalized ER profiles (Fig. S7), which show the similarities in the ER profiles for the pine/fir mixture and spruce. Figure 11 shows the PR results including the 90/10 synthetic mixtures. Both samples were correctly clustered with their respective dominant fuel family. The synthetic mixture results suggest that the algorithm can select marker compounds that can differentiate fuel types even when they are mixed in relatively even proportions (i.e., 60/40 pine/spruce and fir/spruce mixtures) but for some mixtures the differentiation might be poor (i.e., 60/40 pine/fir). Including more mixed fuel samples in the training and test sets can likely improve the separation of complex mixtures achieved using the PR algorithm.

3.3 Classification

~~For the classification algorithm, the scores of the selected PCs were used as input for the LDA training. PC1 and PC4 were selected since they provided better separation across the three fuel types (see section 3.2.3), while explaining 82% of the variability in the data set. The outcome of LDA is a probability, calculated using Eq. 4, that a sample belongs to one of the main clusters determined by k-means (see section 3.2.3).~~[d]

~~$$\log P(y=k|x) = -\frac{1}{2}(x-\mu_k)^t \Sigma^{-1}(x-\mu_k) + logP(y=k) + C_{st} \quad (4)$$~~[d]

~~where k is the class of sample x, μ is the vector of the means for each class based on the selected features and Σ is the common covariance matrix for the three classes in the training set. In this application the probability score is related to the proximity of sample to a class of samples (cluster) in the PCA space (Fig. 10 and Fig. 11) which is linked to its similarity with the emission profiles for the three fuel types (Fig. 4). The assignment of a sample to a class is based on the class with the highest probability,~~

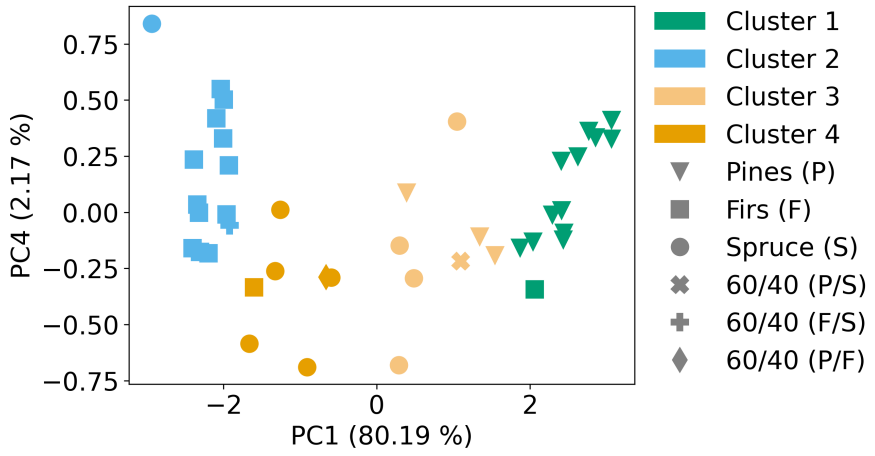

Figure 10. PCA coupled with *k*-means clustering results for the PC1 and PC4 pair including the 60%/40% synthetic mixtures.

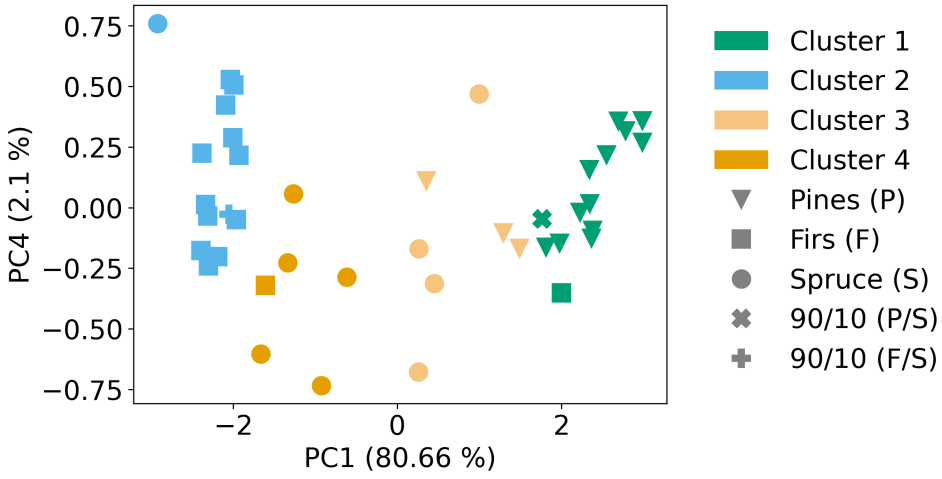

Figure 11. PCA coupled with *k*-means clustering results for the PC1 and PC4 pair including the 90%/10% synthetic mixtures.

~~even if marginally higher. For example a sample with a pine probability score of 70% or more will most likely be inside the pine cluster. Generally, samples with probability scores 60% and higher are most likely in the cluster space of a fuel family. Samples with a probability score 60% and lower are more likely to be adjacent to more than one fuel family in the PCA space. For the training of the classifier all 39 samples from FIREX FL16 were used.~~[d]

3.3.1 Synthetic mixtures and FLAME-4 FL12 samples

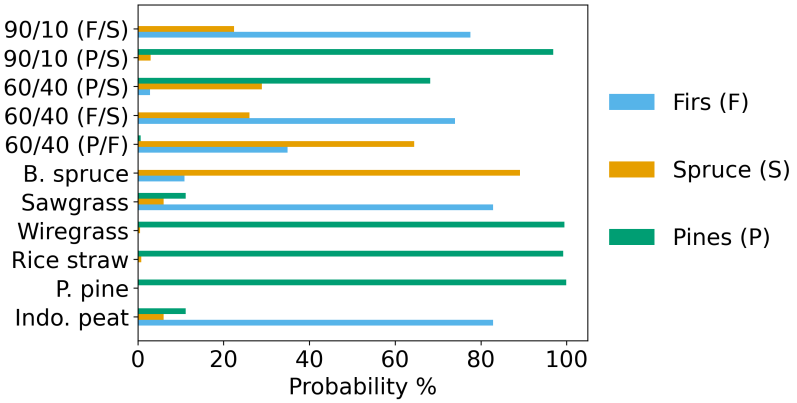

Figure 12. Classification results for synthetic mixtures and FLAME-4 samples.

After the pattern recognition, all 39 samples from the FIREX FL16 data set and the selected compounds in the form of the retained components (PC1 and PC4) were provided to the LDA algorithm for training. As described in section 2.4, LDA provides class membership probability (Eq. 5). In this application the probability score is related to the proximity of a sample to a class of samples (cluster) in the PCA space (Figs. 10 and 11) which is linked to its similarity with the emission profiles for the three fuel types (Fig. 4). The assignment of a sample to a class is based on the class with the highest probability, even if marginally higher. For example a sample with a pine probability score of 70% or more will most likely be inside the pine cluster. Generally, samples with probability scores 60% and higher are most likely in the cluster space of a fuel family. Samples with a probability score of 60% and lower are more likely to be adjacent to one or more fuel families in the PCA space.[n]

The classification algorithm was tested using the synthetic mixtures and FLAME-4 FL12 samples before testing using the BFRS field data. The classification results for the synthetic mixtures are shown in Fig. 12. Two of three 60/40 synthetic mixtures were classified correctly, pine/spruce and fir/spruce, with classification probabilities of 70% and higher for the dominant fuel family. The 60/40 pine/fir synthetic mixture was classified as spruce. Its classification is a result of its clustering in the PCA space with spruce (Fig. 10), which is directly connected to its similarity with the spruce emission profile (Fig. S7). The two 90/10 synthetic mixtures, pine/spruce and fir/spruce, were correctly classified with classification probabilities over 80% for the dominant fuel family. Application of the classifier to the synthetic mixtures demonstrated that the mixtures can be correctly classified based on the dominant fuel family in mixed fuel samples (4 of 5 mixtures); however, incorrect classification can occur when the mixed fuel emission profiles are similar to individual fuel emission profiles, resulting in poorer separation with PCA. The results of the classification algorithm for mixed samples can be improved in future work through expanded testing and training on a broader range of fuel types and relevant mixtures.

The classification results for the FLAME-4 FL12 samples are shown in Fig. 12. This data set included six fuel species (ponderosa pine, black spruce, Indonesian peat, rice straw, wiregrass and sawgrass); only one of which, ponderosa pine, was in the training set. Both ponderosa pine and black spruce samples were classified correctly (Fig. 12), with classification probabilities

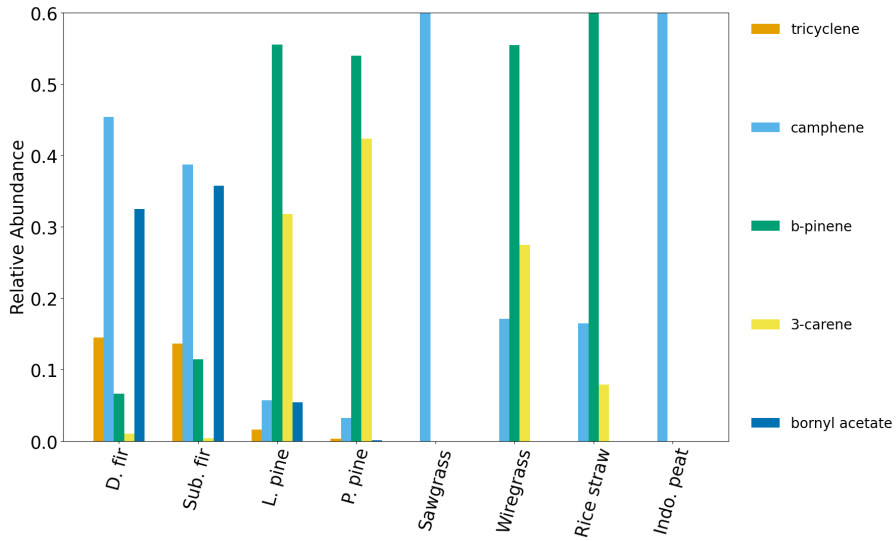

Figure 13. Normalized emission ratio profiles for FIREX FL16 samples: pines and firs; and FLAME-4 FL12 samples: sawgrass, wiregrass, rice straw and Indonesian peat. The relative abundances of camphene and b-pinene in sawgrass, Indonesian peat and rice straw were ≥ 0.6, but for figure clarity, the axis limits were not changed.

over 90%. The Indonesian peat, rice straw, wiregrass and sawgrass samples were classified as firs or pines (Fig. 12), with classification probabilities over 70%. The classification algorithm evaluated partial similarity against only three options (pine, firs or spruce), none of which represent the fuel types of the four fuel species. Figure 13 shows the average normalized emission ratio profiles for pines and firs, as well as Indonesian peat, rice straw, wiregrass and sawgrass. It can be seen that camphene is the only one of the five compounds that is present in the sawgrass and Indonesian peat samples, and thus these fuels were classified as firs, which also have a high relative abundance of camphene. Wiregrass and rice straw samples also include camphene, but have higher relative abundances of b-pinene and 3-carene, and thus were classified as pines, which also have higher relative abundances of these two compounds (Fig. 13). As illustrated by the application to the synthetic fuel mixtures, the performance of the classification algorithm can be improved in future work by expanding the range of fuel types and mixtures included in the training and test sets.

3.3.2 Blodgett samples

Figures 14-16 show the results of the classification algorithm applied to the BFRS samples from three different prescribed burn plots: 60, 340 and 400; the composition of which are shown in Hatch et al. (2019)(SI Fig. S2-S4). Seven different fuel species were identified in the three burned plots: white fir, incense cedar, tanoak, sugar pine, ponderosa pine, Douglas fir and California black oak. Due to the heterogeneity of the fuels, and the influence of meteorology and sampling location, it was not possible to determine the relative contribution of each fuel species to each sample. Instead the average overstory composition (Figs. S8-S11) was used to determine likely influences from dominant sources close to each sampling location. For plot 60,

sites one and two (Fig. S8), the main influence was from firs (47%), followed by similar amounts of pines and incense cedar (25% and 27%), with no contribution from tanoak or California black oak. For plot 60, site three (Fig. S9), the main influence was from incense cedar (43%), followed by firs (34%), pines (26%) and California black oak (10%). For plot 340 (Fig. S10) the main influence was from firs (63%), followed by pines (21%), incense cedar (12%) and (2%) from tanoak and California black oak. Finally for plot 400 (Fig. S11) the main influence was from firs (55%), followed by pines (26%) and incense cedar (18%). The classification algorithm classified all samples from plots 60 (Fig. 14) and 340 (Fig. 15) as fir dominant. Nine out of ten samples from plot 400 (Fig. 16) were classified as fir dominant and one as pine dominant. While spruce was absent in the burned plots, all samples (with the exception of the pine dominant sample in plot 400) had a higher classification probability for spruce than pines.

For plot 60 there were a total of 11 samples collected. Five samples in sites one and two, which is fir dominant (Fig. S8), and six samples in site three, which is incense cedar dominant (Fig. S9). For sites one and two the classifier results were reasonable based on the overstory composition, but for site three the classification results were inconclusive since no emission profiles were available for cedars. It is likely that incense cedar or the mixture of incense cedar with firs most closely resembles the fir emission profiles of the selected compounds and thus was classified as firs. For plots 340 and 400 the classification results, probability of 63% in 340 and 55% in 400, are reasonable given 16 out of 17 samples are fir dominant (Figs. 15-16). The one sample that was classified as pines in plot 400 was most likely affected by ponderosa pine emissions during sampling. The one sample that was classified as pines in plot 400 was most likely affected by ponderosa pine emissions during sampling. The composition plots in Hatch et al. (2019) show that one of the inventoried plots next to plot 400 (sites one and two) had an average fractional overstory composition of more than 50% ponderosa pine. Regarding the elevated probability for spruce, despite its absence from all burned plots, it was likely an artifact of mixed smoke between pines and firs, that was also shown in the synthetic mixed sample PR and classification results. Among the three plots, pines and firs together account for more than 70% of the overstory composition on average. Thus the contribution from both firs and pines could lead to smoke mixtures that resemble the spruce emission profiles. Tanoak and California black oak (Figs. S8-S11) account for 2% - 10% of the total contributions among the three plots. Due to insufficient emission data, their contributions to the smoke samples could not be evaluated, but given their low overstory contribution it is likely that they did not influence the collected samples substantially. The results for the BFRS data showed that the laboratory-based emission profiles selected by the PR algorithm can be applied to smoke samples collected in the field and can be used to identify dominant fuel sources even in mixed smoke samples. While the algorithm has been tested and trained on only three fuel types, widespread application can be achieved with further training and testing using a more diverse set of compounds and broader range of fuel types.

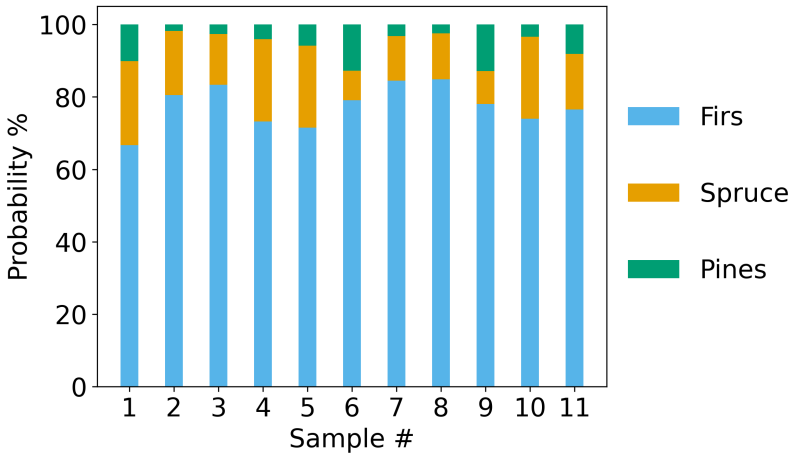

Figure 14. Classification probability by fuel class for plot 60.

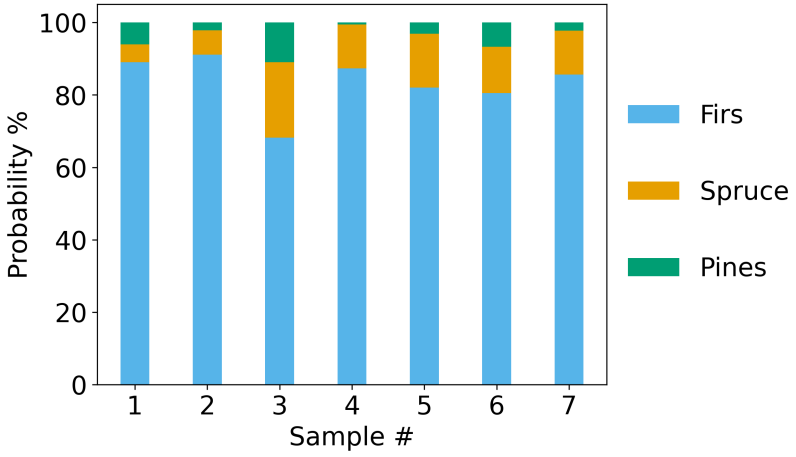

Figure 15. Classification probability by fuel class for plot 340.

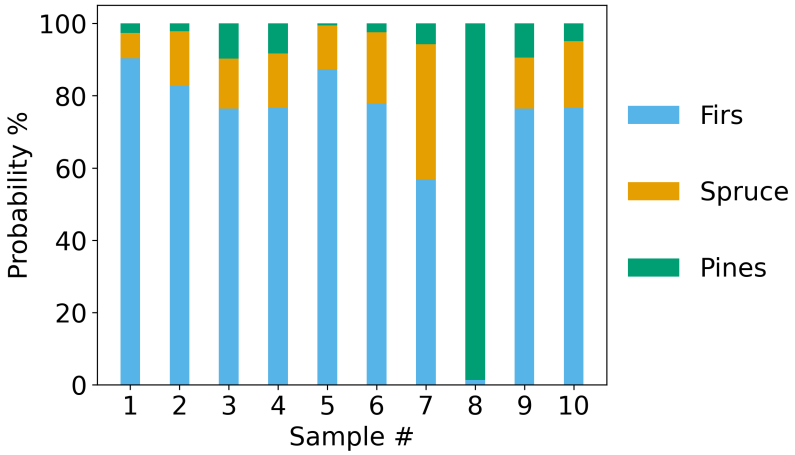

Figure 16. Classification probability by fuel class for plot 400.

4 Conclusions

4.1 ~~Pattern recognition and classification~~[d]

A supervised pattern recognition (PR) algorithm was developed and applied in this study to: 1) differentiate sources/fuel types using NMOCs measured in smoke samples and selected with an ANOVA based feature selection method; and 2) train a classification algorithm to identify dominant sources/fuel types in smoke samples based on the unique speciation profiles identified by the PR algorithm. The PR algorithm was able to group five fuel species (Douglas and subalpine fir, ponderosa and loblolly pine, and Engelmann spruce) into three fuel types (pines, firs and spruce), with minimum overlap; only five of 39 total samples were grouped with types that were not representative of the fuel species. The separation was achieved using five monoterpenoids that the algorithm selected out of a pool of 93. Future work should include exploring how normality and heteroskedasticity, which are underlying assumptions for ANOVA, may be affecting separation. This can be achieved by using non-parametric tests that do not make assumptions about the underlying data distribution and are more robust than ANOVA in the presence of heteroskedasticity.[n] The PR algorithm was tested with five synthetic mixtures, for which three of five (60/40 fir/spruce, 90/10 pine/spruce, and 90/10 fir/spruce) were successfully separated and clustered with their dominant family. The same synthetic mixtures were also tested using the classification algorithm, where four of five were classified correctly (60/40 pine/spruce, 60/40 fir/spruce, 90/10 pine/spruce, and 90/10 fir/spruce). The application of the classification algorithm to the synthetic mixtures demonstrated that dominant source contributions could be identified in fuel mixtures. For the FLAME-4 FL12 samples the classification algorithm correctly classified two of six samples (ponderosa pine and black spruce); these two samples were the only fuels represented by the three fuel types. For the BFRS field samples, based on the fractional overstory composition, the classification results were reasonable with 27 out of 28 samples being classified as fir dominant and one sample as pine dominant. The incorrect classifications that occurred with the synthetic fuel mixture (60/40 pine/fir) and the FLAME-4 FL12 samples (Indonesian peat, rice straw, wiregrass and sawgrass) were due to the similarity or partial similarity of their emission profiles with the fuels used to train the classification model. This can be resolved in future applications by including more compounds and a broader range of fuel types, including in mixtures. This will also facilitate the use of this approach for identifying contributing fuels outside of western coniferous forests.

Code and data availability. The data and implementation scripts for the pattern recognition algorithm and the classification model are available through a GitHub repository https://github.com/christos-stamatis/supervised_pattern_recognition but are not yet final.

Author contributions. CS and KCB contributed equally to the manuscript; CS lead the data analysis, data interpretation and manuscript preparation efforts; KCB lead the conceptual design and manuscript editing.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors want to thank Lindsay Hatch for data acquisition and processing of all FIREX FL16, FLAME-4 FL12 and BFRS samples; Ariel Roughton for the land cover analysis of the BFRS samples; Paul Van Rooy for discussions on the applicability of speciation profiles for data interpretation and smoke modeling; and Vanessa Selimovic for the CO data that were used for the ER calculations.

Financial support. This work was supported by; NOAA grants NA16OAR4310103 and NA17OAR4310007; and CARB grant 00010311.

References

Abdi, H. and Williams, L. J.: Principal component analysis, WIREs Computational Statistics, 2, 433–459, https://doi.org/https://doi.org/10.1002/wics.101, 2010.

Alvarado, M. J. and Prinn, R. G.: Formation of ozone and growth of aerosols in young smoke plumes from biomass burning: 1. Lagrangian parcel studies, Journal of Geophysical Research: Atmospheres, 114, https://doi.org/https://doi.org/10.1029/2008JD011144, 2009.

Andreae, M. O.: Emission of trace gases and aerosols from biomass burning - an updated assessment, ATMOSPHERIC CHEMISTRY AND PHYSICS, 19, 8523–8546, https://doi.org/10.5194/acp-19-8523-2019, 2019.

Andreae, M. O., Browell, E. V., Garstang, M., Gregory, G. L., Harriss, R. C., Hill, G. F., Jacob, D. J., Pereira, M. C., Sachse, G. W., Setzer, A. W., Dias, P. L. S., Talbot, R. W., Torres, A. L., and Wofsy, S. C.: Biomass-burning emissions and associated haze layers over Amazonia, Journal of Geophysical Research: Atmospheres, 93, 1509–1527, https://doi.org/10.1029/JD093iD02p01509, 1988.

Chen, J., Anderson, K., Pavlovic, R., Moran, M. D., Englefield, P., Thompson, D. K., Munoz-Alpizar, R., and Landry, H.: The FireWork v2.0 air quality forecast system with biomass burning emissions from the Canadian Forest Fire Emissions Prediction System v2.03, Geoscientific Model Development, 12, 3283–3310, https://doi.org/10.5194/gmd-12-3283-2019, 2019.

Dennison, P. E., Brewer, S. C., Arnold, J. D., and Moritz, M. A.: Large wildfire trends in the western United States, 1984–2011, Geophysical Research Letters, 41, 2928–2933, https://doi.org/https://doi.org/10.1002/2014GL059576, 2014.

Dong, Y. and Peng, C.-Y. J.: Principled missing data methods for researchers, SpringerPlus, 2, 222, https://doi.org/10.1186/2193-1801-2-222, 2013.

Elkan, C.: Using the Triangle Inequality to Accelerate K-Means, in: Proceedings of the Twentieth International Conference on International Conference on Machine Learning, ICML'03, p. 147–153, AAAI Press, 2003.

Fu, P., Kawamura, K., and Barrie, L. A.: Photochemical and Other Sources of Organic Compounds in the Canadian High Arctic Aerosol Pollution during Winter-Spring, Environmental Science & Technology, 43, 286–292, https://doi.org/10.1021/es803046q, 2009.

Gewers, F. L., Ferreira, G. R., Arruda, H. F. D., Silva, F. N., Comin, C. H., Amancio, D. R., and Costa, L. D. F.: Principal Component Analysis: A Natural Approach to Data Exploration, ACM Comput. Surv., 54, https://doi.org/10.1145/3447755, 2021.

Goode, J. G., Yokelson, R. J., Ward, D. E., Susott, R. A., Babbitt, R. E., Davies, M. A., and Hao, W. M.: Measurements of excess O3, CO2, CO, CH4, C2H4, C2H2, HCN, NO, NH3, HCOOH, CH3COOH, HCHO, and CH3OH in 1997 Alaskan biomass burning plumes by airborne Fourier transform infrared spectroscopy (AFTIR), Journal of Geophysical Research: Atmospheres, 105, 22 147–22 166, https://doi.org/10.1029/2000JD900287, 2000.

Hastie, T., Tibshirani, and Robert Friedman, J.: The Elements of Statistical Learning, Springer, 2009.

Hatch, L. E., Luo, W., Pankow, J. F., Yokelson, R. J., Stockwell, C. E., and Barsanti, K. C.: Identification and quantification of gaseous organic compounds emitted from biomass burning using two-dimensional gas chromatography–time-of-flight mass spectrometry, Atmos. Chem. Phys., 15, 1865–1899, https://doi.org/10.5194/acp-15-1865-2015, 2015.

Hatch, L. E., Yokelson, R. J., Stockwell, C. E., Veres, P. R., Simpson, I. J., Blake, D. R., Orlando, J. J., and Barsanti, K. C.: Multi-instrument comparison and compilation of non-methane organic gas emissions from biomass burning and implications for smoke-derived secondary organic aerosol precursors, Atmospheric Chemistry and Physics, 17, 1471–1489, https://doi.org/10.5194/acp-17-1471-2017, 2017.

Hatch, L. E., Rivas-Ubach, A., Jen, C. N., Lipton, M., Goldstein, A. H., and Barsanti, K. C.: Measurements of I/SVOCs in biomass-burning smoke using solid-phase extraction disks and two-dimensional gas chromatography, Atmos. Chem. Phys., 18, 17 801–17 817, https://doi.org/10.5194/acp-18-17801-2018, 2018.

Hatch, L. E., Jen, C. N., Kreisberg, N. M., Selimovic, V., Yokelson, R. J., Stamatis, C., York, R. A., Foster, D., Stephens, S. L., Goldstein, A. H., and Barsanti, K. C.: Highly Speciated Measurements of Terpenoids Emitted from Laboratory and Mixed-Conifer Forest Prescribed Fires, Environmental Science & Technology, 53, 9418–9428, https://doi.org/10.1021/acs.est.9b02612, pMID: 31318536, 2019.

Holder, A. L., Gullett, B. K., Urbanski, S. P., Elleman, R., O'Neill, S., Tabor, D., Mitchell, W., and Baker, K. R.: Emissions from prescribed burning of agricultural fields in the Pacific Northwest, Atmospheric Environment, 166, 22–33, https://www.sciencedirect.com/science/article/pii/S1352231017304247, 2017.

Hu, Y. Q., Fernandez-Anez, N., Smith, T. E. L., and Rein, G.: Review of emissions from smouldering peat fires and their contribution to regional haze episodes, INTERNATIONAL JOURNAL OF WILDLAND FIRE, 27, 293–312, https://doi.org/10.1071/WF17084, 2018.

Jaffe, D. A., O'Neill, S. M., Larkin, N. K., Holder, A. L., Peterson, D. L., Halofsky, J. E., and Rappold, A. G.: Wildfire and prescribed burning impacts on air quality in the United States, Journal of the Air & Waste Management Association, 70, 583–615, https://doi.org/10.1080/10962247.2020.1749731, pMID: 32240055, 2020.

Jain, A. K.: Data clustering: 50 years beyond K-means, Pattern Recognition Letters, 31, 651–666, https://doi.org/https://doi.org/10.1016/j.patrec.2009.09.011, award winning papers from the 19th International Conference on Pattern Recognition (ICPR), 2010.

Jakobsen, J. C., Gluud, C., Wetterslev, J., and Winkel, P.: When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts, BMC Medical Research Methodology, 17, 162, https://doi.org/10.1186/s12874-017-0442-1, 2017.

Jen, C. N., Liang, Y., Hatch, L. E., Kreisberg, N. M., Stamatis, C., Kristensen, K., Battles, J. J., Stephens, S. L., York, R. A., Barsanti, K. C., and Goldstein, A. H.: High Hydroquinone Emissions from Burning Manzanita, Environmental Science & Technology Letters, 5, 309–314, https://doi.org/10.1021/acs.estlett.8b00222, 2018.

Johnson, K. J. and Synovec, R. E.: Pattern recognition of jet fuels: comprehensive GC×GC with ANOVA-based feature selection and principal component analysis, Chemometrics and Intelligent Laboratory Systems, 60, 225–237, https://doi.org/https://doi.org/10.1016/S0169-7439(01)00198-8, fourth International Conference on Environ metrics and Chemometrics held in Las Vegas, NV, USA, 18-20 September 2000, 2002.

Jolliffe, I.: Principal Component Analysis, Springer, 2002.

Keane, R. E. and Lutes, D.: First-Order Fire Effects Model (FOFEM), pp. 1–5, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-51727-8_74-1, 2018.

Kochanski, A. K., Pardyjak, E. R., Stoll, R., Gowardhan, A., Brown, M. J., and Steenburgh, W. J.: One-Way Coupling of the WRF–QUIC Urban Dispersion Modeling System, Journal of Applied Meteorology and Climatology, 54, 2119 – 2139, https://doi.org/10.1175/JAMC-D-15-0020.1, 2015.

Koss, A. R., Sekimoto, K., Gilman, J. B., Selimovic, V., Coggon, M. M., Zarzana, K. J., Yuan, B., Lerner, B. M., Brown, S. S., Jimenez, J. L., Krechmer, J., Roberts, J. M., Warneke, C., Yokelson, R. J., and de Gouw, J.: Non-methane organic gas emissions from biomass burning: identification, quantification, and emission factors from PTR-ToF during the FIREX 2016 laboratory experiment, Atmos. Chem. Phys., 18, 3299–3319, https://doi.org/10.5194/acp-18-3299-2018, 2018.

Lever, J., Krzywinski, M., and Altman, N.: Principal component analysis, Nature Methods, 14, 641–642, https://doi.org/10.1038/nmeth.4346, 2017.

Lindaas, J., Pollack, I. B., Garofalo, L. A., Pothier, M. A., Farmer, D. K., Kreidenweis, S. M., Campos, T. L., Flocke, F., Weinheimer, A. J., Montzka, D. D., Tyndall, G. S., Palm, B. B., Peng, Q., Thornton, J. A., Permar, W., Wielgasz, C., Hu, L., Ottmar, R. D.,

Restaino, J. C., Hudak, A. T., Ku, I.-T., Zhou, Y., Sive, B. C., Sullivan, A., Collett Jr, J. L., and Fischer, E. V.: Emissions of Reactive Nitrogen From Western U.S. Wildfires During Summer 2018, Journal of Geophysical Research: Atmospheres, 126, e2020JD032 657, https://doi.org/10.1029/2020JD032657, 2021.

Liu, X., Huey, L. G., Yokelson, R. J., Selimovic, V., Simpson, I. J., Müller, M., Jimenez, J. L., Campuzano-Jost, P., Beyersdorf, A. J., Blake, D. R., Butterfield, Z., Choi, Y., Crounse, J. D., Day, D. A., Diskin, G. S., Dubey, M. K., Fortner, E., Hanisco, T. F., Hu, W., King, L. E., Kleinman, L., Meinardi, S., Mikoviny, T., Onasch, T. B., Palm, B. B., Peischl, J., Pollack, I. B., Ryerson, T. B., Sachse, G. W., Sedlacek, A. J., Shilling, J. E., Springston, S., St. Clair, J. M., Tanner, D. J., Teng, A. P., Wennberg, P. O., Wisthaler, A., and Wolfe, G. M.: Airborne measurements of western U.S. wildfire emissions: Comparison with prescribed burning and air quality implications, Journal of Geophysical Research: Atmospheres, 122, 6108–6129, https://doi.org/10.1002/2016JD026315, 2017.

McKenzie, D., O'Neill, S. M., Larkin, N. K., and Norheim, R. A.: Integrating models to predict regional haze from wildland fire, Ecological Modelling, 199, 278–288, https://doi.org/https://doi.org/10.1016/j.ecolmodel.2006.05.029, ecological Models as Decision Tools in the 21st Century, 2006.

McMeeking, G. R., Kreidenweis, S. M., Carrico, C. M., Lee, T., Collett Jr., J. L., and Malm, W. C.: Observations of smoke-influenced aerosol during the Yosemite Aerosol Characterization Study: Size distributions and chemical composition, Journal of Geophysical Research: Atmospheres, 110, https://doi.org/https://doi.org/10.1029/2004JD005389, 2005.

McNeish, D.: Missing data methods for arbitrary missingness with small samples, Journal of Applied Statistics, 44, 24–39, https://doi.org/10.1080/02664763.2016.1158246, 2017.

Miller, J. D., Safford, H. D., Crimmins, M., and Thode, A. E.: Quantitative Evidence for Increasing Forest Fire Severity in the Sierra Nevada and Southern Cascade Mountains, California and Nevada, USA, Ecosystems, 12, 16–32, https://doi.org/10.1007/s10021-008-9201-9, 2009.

Nelson, K. J., Connot, J., Peterson, B., and Martin, C.: The LANDFIRE Refresh Strategy: Updating the National Dataset, Fire Ecology, 9, 80–101, https://doi.org/10.4996/fireecology.0902080, 2013.

Ottmar, R.: Consume 3.0—A Software Tool for Computing Fuel Consumption, Fire Science Brief, p. 6, https://www.firescience.gov/projects/briefs/98-1-9-06_FSBrief55.pdf, 2009.

Park, R. J., Jacob, D. J., Kumar, N., and Yantosca, R. M.: Regional visibility statistics in the United States: Natural and transboundary pollution influences, and implications for the Regional Haze Rule, Atmospheric Environment, 40, 5405–5423, https://www.sciencedirect.com/science/article/pii/S1352231006004316, 2006.

Pavlovic, R., Chen, J., Anderson, K., Moran, M. D., Beaulieu, P.-A., Davignon, D., and Cousineau, S.: The FireWork air quality forecast system with near-real-time biomass burning emissions: Recent developments and evaluation of performance for the 2015 North American wildfire season, Journal of the Air & Waste Management Association, 66, 819–841, https://doi.org/10.1080/10962247.2016.1158214, pMID: 26934496, 2016.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Permar, W., Wang, Q., Selimovic, V., Wielgasz, C., Yokelson, R. J., Hornbrook, R. S., Hills, A. J., Apel, E. C., Ku, I.-T., Zhou, Y., Sive, B. C., Sullivan, A. P., Collett Jr, J. L., Campos, T. L., Palm, B. B., Peng, Q., Thornton, J. A., Garofalo, L. A., Farmer, D. K., Kreidenweis, S. M., Levin, E. J. T., DeMott, P. J., Flocke, F., Fischer, E. V., and Hu, L.: Emissions of Trace Organic Gases From West-

ern U.S. Wildfires Based on WE-CAN Aircraft Measurements, Journal of Geophysical Research: Atmospheres, 126, e2020JD033 838, https://doi.org/10.1029/2020JD033838, 2021.

Prichard, S., Larkin, N. S., Ottmar, R., French, N. H., Baker, K., Brown, T., Clements, C., Dickinson, M., Hudak, A., Kochanski, A., Linn, R., Liu, Y., Potter, B., Mell, W., Tanzer, D., Urbanski, S., and Watts, A.: The Fire and Smoke Model Evaluation Experiment—A Plan for Integrated, Large Fire–Atmosphere Field Campaigns, Atmosphere, 10, https://doi.org/10.3390/atmos10020066, 2019.

Prichard, S. J., O'Neill, S. M., Eagle, P., Andreu, A. G., Drye, B., Dubowy, J., Urbanski, S., and Strand, T. M.: Wildland fire emission factors in North America: synthesis of existing data, measurement needs and management applications, International Journal of Wildland Fire, 29, 132–147, https://doi.org/10.1071/WF19066, 2020.

Reeves, M. C., Ryan, K. C., Rollins, M. G., and Thompson, T. G.: Spatial fuel data products of the LANDFIRE Project, International Journal of Wildland Fire, 18, 250–267, https://doi.org/10.1071/WF08086, 2009.

Sekimoto, K., Koss, A. R., Gilman, J. B., Selimovic, V., Coggon, M. M., Zarzana, K. J., Yuan, B., Lerner, B. M., Brown, S. S., Warneke, C., Yokelson, R. J., Roberts, J. M., and de Gouw, J.: High- and low-temperature pyrolysis profiles describe volatile organic compound emissions from western US wildfire fuels, Atmos. Chem. Phys., 18, 9263–9281, https://doi.org/10.5194/acp-18-9263-2018, 2018.

Selimovic, V., Yokelson, R. J., Warneke, C., Roberts, J. M., de Gouw, J., Reardon, J., and Griffith, D. W. T.: Aerosol optical properties and trace gas emissions by PAX and OP-FTIR for laboratory-simulated western US wildfires during FIREX, Atmos. Chem. Phys., 18, 2929–2948, https://doi.org/10.5194/acp-18-2929-2018, 2018.

Simoneit, B. R.: Biomass burning — a review of organic tracers for smoke from incomplete combustion, Applied Geochemistry, 17, 129–162, https://www.sciencedirect.com/science/article/pii/S0883292701000610, 2002.

Simpson, I. J., Akagi, S. K., Barletta, B., Blake, N. J., Choi, Y., Diskin, G. S., Fried, A., Fuelberg, H. E., Meinardi, S., Rowland, F. S., Vay, S. A., Weinheimer, A. J., Wennberg, P. O., Wiebring, P., Wisthaler, A., Yang, M., Yokelson, R. J., and Blake, D. R.: Boreal forest fire emissions in fresh Canadian smoke plumes: C-1-C-10 volatile organic compounds (VOCs), CO2, CO, NO2, NO, HCN and CH3CN, ATMOSPHERIC CHEMISTRY AND PHYSICS, 11, 6445–6463, https://doi.org/10.5194/acp-11-6445-2011, 2011.

Stockwell, C. E., Yokelson, R. J., Kreidenweis, S. M., Robinson, A. L., DeMott, P. J., Sullivan, R. C., Reardon, J., Ryan, K. C., Griffith, D. W. T., and Stevens, L.: Trace gas emissions from combustion of peat, crop residue, domestic biofuels, grasses, and other fuels: configuration and Fourier transform infrared (FTIR) component of the fourth Fire Lab at Missoula Experiment (FLAME-4), Atmospheric Chemistry and Physics, 14, 9727–9754, https://doi.org/10.5194/acp-14-9727-2014, 2014.

Stockwell, C. E., Veres, P. R., Williams, J., and Yokelson, R. J.: Characterization of biomass burning emissions from cooking fires, peat, crop residue, and other fuels with high-resolution proton-transfer-reaction time-of-flight mass spectrometry, Atmospheric Chemistry and Physics, 15, 845–865, https://doi.org/10.5194/acp-15-845-2015, 2015.

Urbanski, S.: Wildland fire emissions, carbon, and climate: Emission factors, Forest Ecology and Management, 317, 51–60, https://www.sciencedirect.com/science/article/pii/S0378112713003447, 2014.

Urbanski, S. P.: Combustion efficiency and emission factors for wildfire-season fires in mixed conifer forests of the northern Rocky Mountains, US, Atmos. Chem. Phys., 13, 7241–7262, https://doi.org/10.5194/acp-13-7241-2013, 2013.

Urbanski, S. P., Hao, W. M., and Baker, S.: Chapter 4 Chemical Composition of Wildland Fire Emissions, vol. 8 of *Wildland Fires and Air Pollution*, pp. 79–107, Elsevier, https://www.sciencedirect.com/science/article/pii/S1474817708000041, 2008.

Vogelmann, J. E., Kost, J. R., Tolk, B., Howard, S., Short, K., Chen, X., Huang, C., Pabst, K., and Rollins, M. G.: Monitoring Landscape Change for LANDFIRE Using Multi-Temporal Satellite Imagery and Ancillary Data, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 4, 252–264, https://doi.org/10.1109/JSTARS.2010.2044478, 2011.

Wan, X., Kawamura, K., Ram, K., Kang, S., Loewen, M., Gao, S., Wu, G., Fu, P., Zhang, Y., Bhattarai, H., and Cong, Z.: Aromatic acids as biomass-burning tracers in atmospheric aerosols and ice cores: A review, Environmental Pollution, 247, 216–228, https://www.sciencedirect.com/science/article/pii/S026974911834644X, 2019.

Ward, D. E. and Hardy, C. C.: Smoke emissions from wildland fires, Environment International, 17, 117–134, https://www.sciencedirect.com/science/article/pii/0160412091900958, 1991.

Welke, J. E., Manfroi, V., Zanus, M., Lazzarotto, M., and Alcaraz Zini, C.: Differentiation of wines according to grape variety using multivariate analysis of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection data, Food Chemistry, 141, 3897–3905, https://doi.org/https://doi.org/10.1016/j.foodchem.2013.06.100, 2013.

Westerling, A. L., Hidalgo, H. G., Cayan, D. R., and Swetnam, T. W.: Warming and Earlier Spring Increase Western U.S. Forest Wildfire Activity, Science, 313, 940–943, https://doi.org/10.1126/science.1128834, 2006.

Yokelson, R. J., Goode, J. G., Ward, D. E., Susott, R. A., Babbitt, R. E., Wade, D. D., Bertschi, I., Griffith, D. W. T., and Hao, W. M.: Emissions of formaldehyde, acetic acid, methanol, and other trace gases from biomass fires in North Carolina measured by airborne Fourier transform infrared spectroscopy, Journal of Geophysical Research: Atmospheres, 104, 30 109–30 125, https://doi.org/https://doi.org/10.1029/1999JD900817, 1999.

Yokelson, R. J., Burling, I. R., Gilman, J. B., Warneke, C., Stockwell, C. E., de Gouw, J., Akagi, S. K., Urbanski, S. P., Veres, P., Roberts, J. M., Kuster, W. C., Reardon, J., Griffith, D. W. T., Johnson, T. J., Hosseini, S., Miller, J. W., Cocker III, D. R., Jung, H., and Weise, D. R.: Coupling field and laboratory measurements to estimate the emission factors of identified and unidentified trace gases for prescribed fires, Atmos. Chem. Phys., 13, 89–116, https://doi.org/10.5194/acp-13-89-2013, 2013.

Zangrando, R., Barbaro, E., Zennaro, P., Rossi, S., Kehrwald, N. M., Gabrieli, J., Barbante, C., and Gambaro, A.: Molecular Markers of Biomass Burning in Arctic Aerosols, Environmental Science & Technology, 47, 8565–8574, https://doi.org/10.1021/es400125r, 2013.

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Allan, J. D., Coe, H., Ulbrich, I., Alfarra, M. R., Takami, A., Middlebrook, A. M., Sun, Y. L., Dzepina, K., Dunlea, E., Docherty, K., DeCarlo, P. F., Salcedo, D., Onasch, T., Jayne, J. T., Miyoshi, T., Shimono, A., Hatakeyama, S., Takegawa, N., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Williams, P., Bower, K., Bahreini, R., Cottrell, L., Griffin, R. J., Rautiainen, J., Sun, J. Y., Zhang, Y. M., and Worsnop, D. R.: Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere midlatitudes, Geophysical Research Letters, 34, https://doi.org/10.1029/2007GL029979, 2007.

Zhang, Y., Kong, S., Sheng, J., Zhao, D., Ding, D., Yao, L., Zheng, H., Wu, J., Cheng, Y., Yan, Q., Niu, Z., Zheng, S., Wu, F., Yan, Y., Liu, D., and Qi, S.: Real-time emission and stage-dependent emission factors/ratios of specific volatile organic compounds from residential biomass combustion in China, Atmospheric Research, 248, 105 189, https://www.sciencedirect.com/science/article/pii/S016980952031125X, 2021.

Ziółkowska, A., Wąsowicz, E., and Jeleń, H. H.: Differentiation of wines according to grape variety and geographical origin based on volatiles profiling using SPME-MS and SPME-GC/MS methods, Food Chemistry, 213, 714–720, https://doi.org/https://doi.org/10.1016/j.foodchem.2016.06.120, 2016.